International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Text and Data Formatting for Machine Learning

Balika J. Chelliah, Arth Jain, Utkarsh Singh, Garima Mehta

***Abstract:** Machine learning is a prominent tool for getting data from large amounts of information. Whereas a good amount of machine learning analysis has targeted on increasing the accuracy and potency of coaching and reasoning algorithms, there is less attention within the equally vital issues of observing the standard of information fed into the machine learning model. The standard of huge information is far away from good. Recent studies have shown that poor quality will bring serious errors to the result of big data analysis and this could have an effect on in making additional precise results from the information. Advantages of data preprocessing within the context of ML are advanced detection of errors, model-quality improves by the usage of better data, savings in engineering hours to debug issues*

***Keywords:** Data Science ,Dataset ,Text Preprocessing*

I. INTRODUCTION

There is little doubt that the industries are going ablaze with the massive eruption of data. None of the sectors have remained untouched of this forceful modification during a decade. [3]Technology has crept within every business arena and therefore, it's become a vital a part of each and every process unit. Talking regarding IT trade specifically, code and automation are the minimum essential terms and are utilized in every and each section of a process and method cycle.

Businesses are focusing a lot on dexterity and innovation instead of stability and adopting the technologies of big data to facilitate the businesses accomplish that in no time.[1]Big Data analytics has not solely allowed the companies to remain updated with the dynamical dynamics but also has allowed them to predict the long-term trends giving a competitive edge. The main centre of interest was on building framework and solutions to stock data. Currently once Hadoop and different frameworks have with success resolved the matter of storage, the main focus has shifted to the processing of this data. Data Science is the charm here.

All the ideas that are seen in Hollywood sci-fi movies will really grow to be reality by Data Science. Data Science is the future of AI.

[8]Data Science is a set of tools that we tend to use to explore, clean and model knowledge so as to extract real-world, purposeful data. Obtaining real-world data first needs real-world data — that real-world data is faulty and dirty.

The collection of data by corporations massive and tiny typically done by an amateur; sometimes somebody at the corporate however typically it’s a consumer of a user of the company’s product. They’ll insert the data via a user-interface where several of the fields are non-obligatory. The user will enter in their data in many alternative designs and formats too. All of this ends up in dirty data.

Before being able to run the data through a Machine Learning model, it has to be compelled to clean it up a little. [4]Data clean-up is one of those things that everybody will do however nobody talks regarding to it. There aren’t any hidden tricks and secrets to uncover. However, correct data clean-up will build or break your project. Skilled data scientists typically invest an awfully giant portion of their time on this step.

If there is a properly clean dataset, even straightforward algorithms will learn spectacular insights from the data. Clearly, different kinds of data would need different kinds of clean-up.The future of data processing lies within the cloud. [2]Cloud technology builds on the convenience of current electronic data processing ways and accelerates its speed and effectiveness. [5]Faster, higher-quality data suggests that a lot of data for every organization to utilize and a lot of valuable insights to extract.

Big Data is altering how everyone does business. Today, remaining agile and competitive depends on having an effective and clear data processing strategy. [5]The cloud has driven enormous advances in technology that deliver the foremost advanced, efficient, and quickest data processing ways to this date.

To make multileveled equations, it should be necessary to treat the equation as a graphic image and insert it into the text when your paper is titled.

[1]In any machine learning task, cleaning or preprocessing the data is as important as model building if not more. And when it comes to unstructured data like text, this process is even more important.

Some of the common text preprocessing / cleaning steps are:

- Removing of Punctuations
- Removing of Stopwords
- Lemmatization
- Removal of emoticons
- Conversion of emojis to words
- Removal of URLs

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author

Dr Balika J.Chelliah*, Department of Computer Science Engineering, SRM Institute of Science and Technlogy, Chennai, India. Email: balika888@gmail.com

Arth jain, Department of Computer Science Engineering, SRM Institute of Science and Technology,Ramapuram, Chennai, India. Email: arthjainabc@gmail.com

Utkarsh singh, Department of Computer Science Engineering.SRM Institute of Science and Technology Ramapuram, Chennai, India. Email:utkarshsinghgr88@gmail.com

Garima Mehta, Department of Computer Science and Technology, SRM Institute of Science and Technology, Chennai, India.

Retrieval Number: A5216119119/2019©BEIESP
DOI: 10.35940/ijitee.A5216.119119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

- Removal of HTML tags
- Chat words conversion
- Spelling correction
- Conversion of emoticons to words
- Stemming
- Converting to Lower case
- Removing of Frequent words

These are the different types of text preprocessing steps which we can do on text data. But we need not do all of these all the times. We need to carefully choose the preprocessing steps based on our use case since that also play an important role.

II. RELATED WORKS

The data set was downloaded online from Kaggle. The research starts by understanding and studying the various processes to clean up the raw data set. There are many available methodologies to do data and text formatting. All these methods do not provide the same result or the results with the same accuracy. For the Machine Learning model to work as efficiently as possible these data anomalies should be minimized to a very large extent. But even after the data clean up there as still some anomalies left that cannot be removed properly. These are the reasons behind the errors and the low accuracy of the Machine Learning models. In this paper a slightly different approach is taken to ensure the accuracy of the model is more. In many approaches the columns or rows with the missing values are taken into consideration, but in this paper the missing values are replaced by mean or median values depending upon the data set. Also, the columns that contain too much missing values are dropped and are not used to make the model so as to prevent the accuracy of the model from going down. The main aim here is to make a data set as clean as possible and attain the maximum accuracy in the Machine Learning models.

III. METHODOLOGY & IMPLEMENTATION

1) Data Collection

This is the most essential starting step because it addresses common challenges, including:

• Automatically deciding relevant attributes in a data string kept in a .csv (comma-separated) file

• Parsing highly-nested data structures like those from XML or JSON files into a table form, for easier scanning and pattern detection.

• Searching and distinguishing relevant data from external repositories.

When considering a DP answer, confirm that it will merge multiple files into one input, like once you have a group of files representing daily transactions, however your machine learning model must ingest a year of data. Also, make certain to possess a contingency set up in place for overcoming issues related to sampling and bias in your data set and your machine learning model.

2) Data Preparation

A) Data Cleaning:

Data is cleaned through processes like filling in missing values, smoothing the blatant data, orsorting out the inconsistencies within the data. Since missing values will tangibly scale back prediction accuracy, form this issue a priority.

In terms of machine learning, supposed or approximated values are more favourable for an algorithm rather than the missing ones. [7] If the person don't apprehend the precise value, ways exist to finely assume the value that is missing or bypass the difficulty altogether. Selecting the correct approach too heavily depends on data and therefore the domain the person has got. Substitute missing values with dummy values, e.g. n/a for categorical or zero for numerical values Substitute the missing numerical values with mean figures. For categorical values, you may use the most frequent data to fill in.

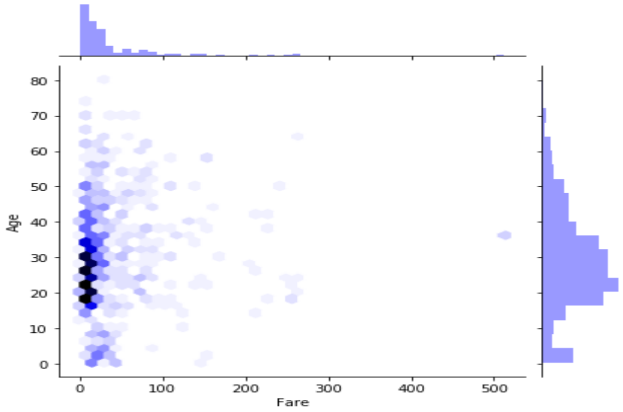

Fig 4.1 The Outliers in the dataset

For example, if a data set is taken that consists of data about Passenger ID, number of cabins and the ages of the passengers and represent it on a graph. It can be seen in Fig 4.2 that roughly 20 percent of the Age data is missing. The proportion of Cabin missing is likely small enough for reasonable replacement with some form of imputation. At the Age column, it looks like there is just missing too much of that data to do something useful with at a basic level and to fill those value.

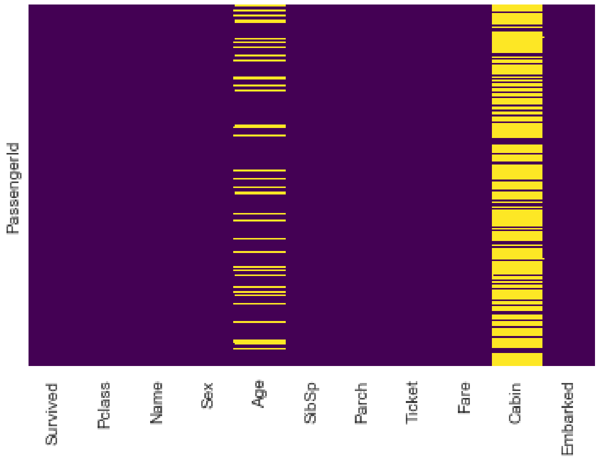

Fig 4.2 Plot with missing data

The Rows from Age column are dropped in Fig 4.3, as too much data is missing and the Age data will not be useful in the further steps because of this reason. By dropping the useless data, the data is made more effective for analysis.

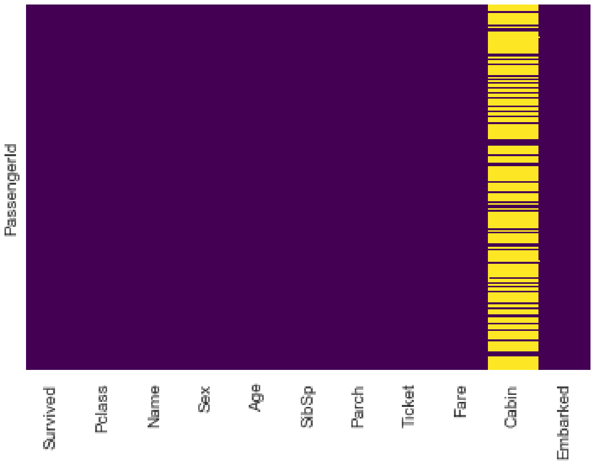

Fig 4.3 Plot without missing data

B) Data Integration:

Data integration is the method of merging data from totally different sources into one, unified read. Integration begins with the ingestion process, and includes steps like cleansing, ETL mapping, and transformation.
Data integration ultimately permits analytic tools so as to supply effective, applicative business intelligence
There is no universal approach or way to data integration. Data integrationsolutions generally involve some
common components, as well as a network of informatio or data sources, a master server, and users and clients accessing data and information from the master server.
In a typical data integration method, the user sends a request to the master server for data. [3] The master server then intakes the required data from internal and external sources. The information and the data is extracted from the sources, then consolidated into one, cohesive data set. This is often served back to the user to be used.

C) Data Transformation:

Data transformation is the method of changing information or data from one format to a different format, typically from the format of a source system into the desired format of a new destination system. [1]The typical method involves changing documents; however, data conversions typically involve the conversion of a program from one computer-oriented language to a different one to allow the program to run on a different platform.
In real world, data transformation involves the employment of a special program that is ready to browse the data’s original base language, verify the language into the data that has to be translated for it to be usable by the new program or system, then begin to rework that data.
Data Transformation consists two key stages:
Data Mapping: The assignment of components from the source base or system toward the destination to capture all altercations that takes place. This is made additionally complex once there advanced transformations like many-to one or one-to-many rules for transformation Code Generation: The creation of the particular actual transformation program. The ensuing data map specifications is employed to form associate practicable program to run on laptop systems.
Commonly used transformational languages:
• Perl: A high-level procedural and object-oriented language capable of powerful computations and operations
• AWK: one amongst the oldest languages and a well liked TXT transformation language
•XSLT: It is an XML data transformation language
• TXL: A prototyping language largely used for source code and ASCII text file transformation
Template Languages and Processors: These specialize in data- to-document transformation

D) Data Reduction:

A database or data warehouse might store terabytes of information or data. So, it's going to take terribly long to perform data analysis and mining on such monstrous amounts of data. Data reduction techniques will be applied to get a reduced illustration of the data set that's abundantly smaller in volume however still contain essential information.
Data Reduction Strategies as seen in Fig 4.4 : -
Data Cube Aggregation:
Aggregation operations are done on the data within the construction of a data cube.
Dimensionality Reduction:
In dimensional reduction useless attributes are detected and removed that cut back the data set size.
Data Compression:
Encoding mechanisms are utilized to cut back the size of the data set.
Numerosity Reduction:
This technique replaces the original data by compressed form of data representation.

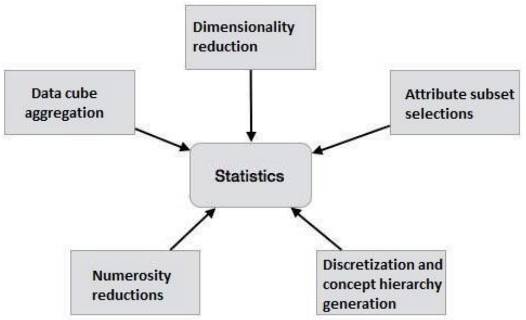

Fig 4.4 Data Reduction

3) Data Input

The final step is to separate your data into two sets; one for coaching your algorithmic program, and another for analysis functions.

Retrieval Number: A5216119119/2019©BEIESP
DOI: 10.35940/ijitee.A5216.119119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Take care to pick out non-overlapping subsets of your data for the coaching and analysis sets so as to make sure correct testing. [9]Invest in tools that offer versioning and cataloguing of your original supply along your already prepared data for input to machine learning algorithms, and also the lineage between them. This way, you'll trace the end result of your predictions back to the input data to refine and optimize your models over time.

IV. RESULT AND DISCUSSION

Data improvement may be a essential method for the success of any machine learning performance. For creating any machine learning model approx. 70 percent of the time spent on data cleaning There are varied alternative strategies of processing your dataset and creating your dataset error-proof. The cleaning of the data must be done in a certain manner so that processed dataset can be used for predictive modelling. While doing sentimental analysis the statement it plays a vital role in the performance of the model is for performing analysis the dataset must and thus providing smoothness to the Model Results from the applied model in Machine Learning shows the format of the data should be in a correct manner. Some particular Machine Learning model wants info in a very specified format, for instance, Random Forest algorithmic program doesn't support null values, thus to execute random forest algorithmic program null values ought to be managed from source raw dataset Another feature is that data set ought to be formatted in such a way that one Machine Learning and Deep Learning algorithms are performed in one data set, and best out of them is chosen

V. CONCLUSION

Treating the data cleaning issue as a large-scale machine learning problem is high-principled and a clear way to address these problems. A principled probabilistic framework will function as the required “melting pot” for all the signals of errors, and therefore the correlation and dependencies among these signals.

Repairing records is principally regarding “predicting” the right values of incorrect or missing attributes within a source data set. The secret is to create models that take into consideration all doable “signals” in predicting the missing or incorrect values. [4]As an example, within the HoloClean system, every cell present in the source table may be a random variable. Dictionaries, reference datasets, accessible quality rules, trustworthy parts of the data are all leveraged to predict the “maximum likelihood” values of these random variables.

The HoloClean model addresses the “holistic cleaning” issue well from a modelling and illustration perspective, permitting these signals that is being considered in isolation to act and guide the process of data prediction. [3]The engineering challenges stay and embody scale, managing and generating training data and involving users in successful manner. These are the challenges every professional faces that the community of data management research well is aware of and is well-suited to resolve.

REFERENCES

1. Data Mining Library for Big Data Processing Platforms: A Case Study-Sparkling Water Platform Elif Cansu Yıldız ; Mehmet S. Aktas ; Oya Kalıpsız ; Alper Nebi Kanlı ; Umut Orçun Turgut 2018 3rd International Conference on Computer Science and Engineering (UBMK) Year: 2018 | Conference Paper | Publisher: IEEE
2. A Survey on Big Data Pre-processing Zhibin Guan ; Tongkai Ji ; Xu Qian ; Yan Ma ; Xuehai Hong 2017 5th Intl Conf on Applied Computing and Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSII-BCD) Year: 2017 | Conference Paper | Publisher: IEEE
3. Design of a scalable data stream channel for big data processing Yong-Ju Lee ; Myungcheol Lee ; Mi-Young Lee ; Sung Jin Hur ; Okgee Min 2015 17th International Conference on Advanced Communication Technology (ICACT) Year: 2015 | Conference Paper | Publisher: IEEE
4. Near Real-Time Big-Data Processing for Data Driven Applications Jānis Kampars ; Jānis Grabis 2017 International Conference on Big Data Innovations and Applications (Innovate-Data) Year: 2017 | Conference Paper | Publisher: IEEE
5. Financial Data Mining Based on Support Vector Machines and Ensemble Learning Shi Lei ; Ma Xinming ; Xi Lei ; Hu Xiaohong 2010 International Conference on Intelligent Computation Technology and Automation Year: 2010 | Volume: 2 | Conference Paper | Publisher: IEEE
6. Big Data Analysis Using Hadoop Framework and Machine Learning as Decision Support System (DSS) (Case Study: Knowledge of Islam Mindset) Nurhayati ; Busman ; Victor Amrizal 2018 6th International Conference on Cyber and IT Service Management (CITSM) Year: 2018 | Conference Paper | Publisher: IEEE
7. Digital Data Forgetting: A Machine Learning Approach Melike Günay ; Eyyüp Yildiz ; Yağiz Nalcakan ; Batuhan Aşiroĝlu ; Ahmet Zencírlí ; Büşra Rümeysa Mete ; Tolga Ensarí 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) Year 2018 | Conference Paper | Publisher: IEEE
8. Improving ML Training Data with Gold-Standard Quality Metric Leslie Barrett, Michael W. Sherman KDD '19: 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Workshop on Data Collection, Curation, and Labeling (DCCL) for Mining and Learning, August 05, 2019, Anchorage, AK. (to appear
9. Building Large Machine Reading-Comprehension Datasets using Paragraph Vectors Radu Soricut, Nan Ding Arxiv, https://arxiv.org/abs/1612.04342 (2016)
10. Debiasing Embeddings for Fairer Text Classification Flavien Prost, Nithum Thain, Tolga Bolukbasi 1st ACL Workshop on Gender Bias for Natural Language Processing (2019)
11. Text Classification with Few Examples using Controlled Generalization Abhijit Mahabal, Jason Baldridge, Burcu Karagol Ayan, Vincent Perot, Dan Roth Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics

AUTHORS PROFILE

Arth jain is currently pursueing Btech in computer Science at SRM Institute of Science and Technology Ramapuram. He has strong interest in Machin Learning and Data Science.He is an aspiring data Scientist.
Email:-arthjainabc@gmail.com

Utkarsh Singh is currently pursuing his Btech in Computer Science and Engineering at SRM Institute of Science and Technology, Chennai. He is interested in Machine Learning and Data Science and plans on pursuing his higher studies in Artificial Intelligence.
Email:-utkarshsinghgr88@gmail.com

Garima Mehta She is undergrad at SRM Institute of Science and technology in Computer Science and technology. She is a Web Developer and She is pursuing it as her future. **Email**:-mehtagarima166@gmail.com

DR. Balika J Chelliah is an Associate Professor in Department of Computer Science & Engineering, SRM Institute of Science and Technology, Chennai. He received his Master and Ph.D degrees in Computer Science & Engineering from SRM Institute Of Technology **Email**:-balika888@gmail.com